Relevancy Ranking of Data Collections

Executive Summary

Relevancy ranking of data collections in EOSDIS search engines is in general suboptimal and may even be largely responsible for users' dissatisfaction with EOSDIS data discovery. However, a number of fairly simple heuristics could be applied to improve that ranking dramatically. Furthermore, it may be possible to recruit relevancy ranking developers from both within and without EOSDIS. Therefore, I recommend that the construction of CMR include mechanisms that would enable relevancy ranking developers to develop and eventually supply algorithms to EOSDIS to improve our relevancy ranking and overall data collection discovery.

Data Discovery, Why So Difficult?

Whenever the number of results returned exceeds what the user can easily scan (say, 7), relevancy ranking becomes critical, especially when results go below the fold or into the next page. This may be one of the key reasons that data search interfaces in EOSDIS are underperforming in the eyes of users (e.g., ACSI surveys). Document-based and keyword-frequency ranking for free text searches, while inexpensive, do not do well in ranking data collections for the likely relevance to the data user.

This spreadsheet documents an experiment in searching for ozone data collections at the GES DISC through three EOSDIS search engines: Mirador, ECHO and GCMD (executed via the OpenSearch interfaces). Alarmingly, they:

1. return different numbers of results
2. return the results in quite different orders
3. return the results in suboptimal order (all of them)

With such discrepancies, it is little wonder that users, particularly new users of EOSDIS, have so much difficulty in finding data collections of relevance to them.

Relevancy Heuristics

However, the homogeneity of the results (they are all data collections) and our knowledge of what most users want allow us to provide much better relevancy ranking using some basic heuristics, proposed as follows. Note that many of them are stated as comparisons, with an eye toward using them in a sort function. They are also listed in a proposed order of application.

Data Content Search-Keyword Heuristic

If a search keyword describes the data content (i.e., the measurement in the data), such as "ozone" that result is more relevant than if it appears elsewhere, such as in the instrument name.

e.g.: search=Ozone: "OMI/Aura Ozone (O3) Total Column 1-Orbit Level 2 Swath 13x24 Km" > "OMI/Aura Sulphur Dioxide (SO2) Total Column 1-Orbit Level 2 Swath 13x24 Km", where OMI = Ozone Monitoring Instrument.

Data Content Modifier Corollary

If a search keyword represents the "Noun" in the data content, such as Ozone, this is more useful in the above heuristic than if it is a modifier. Thus "Column", as in "Total Column Ozone" or "Total Column SO2" is less useful than "Ozone" in this respect. The curated GCMD keywords, which focus more on the "nouns" than the modifiers may be useful in discriminating these two.

Temporal and Spatial Coverage Heuristic

When the user has supplied a date range, data collections that cover the entire range are more relevant than data collections that cover only part of the range.

When the user has supplied a spatial constraint, data collections that cover the entire constraint are more relevant than data collections that cover only part of the constraint.

Data Version Heuristic

Two versions of the same data collection should have adjacent relevancies, with the later version as the more relevant. (This may be overridden by the Data Content Search-Keyword Heuristic when measurements are added to later versions.)

Ease of Use Heuristic

For several categories of users, such as less experienced or educational, (see User Intent Modeling below), Ease of Use can be an important relevance factor. In general, processing level can be used as a rough proxy for this:

Level 0 is of use only to science teams

Level 1 to users who have geophysical retrieval algorithms

Level 2 requires quality screening and geolocation of swaths

Level 3 is gridded and therefore amenable to many off-the-shelf tools

Level 4 is not only gridded, but usually free of gaps

(This is not a strict ranking, as Level 1 is used for RGB images by many users, while many Level 4 parameters require a high degree of expertise to understand. Also, there may be gradations within processing levels, with, say, cylindrical projections being easier to use than sinusoidal Level 3 products.)

Novelty Heuristic

NASA has an interest in giving newly released data collections enhanced visibility. This should be applied before the Popularity and Impact Metric heuristics (below).

Data Popularity Heuristic

A datasets that has been downloaded by more users than another dataset should have higher relevancy ranking in the search results.

Impact Metric Heuristic

An alternative to the Data Popularity is the Impact Metric Heuristic. However, this likely needs more years of data citation to compute reliable metrics for these.

User Intent Modeling

Many search providers apply User Intent Modeling to predict and classify search behavior, thus helping them to optimize search results. While formal modeling approaches are likely beyond the scope of what EOSDIS search providers can apply (but see below on Implementation Approach), we can again formulate some heuristics that may improve relevancy ranking for user. Some of these heuristics are inferring the user's level of experience and expertise in remote sensing data while others base the inference on what the user is most likely trying to apply the data to. Useful information can come from the following locations:

1. the search words themselves
2. referring portals and applications
3. user registration profile

User Interests Heuristic

The advent of user registration provides a golden opportunity to tune searches to a user's expressed interests. To enable this would require the ability for users to express those interests in the User Registration profile. Note that these are NOT the same as the current user type designation. Rather, this would provide a list of research areas (e.g., following the AGU section breakdown), specific types of applications (e.g., Landslides) and even specific educational areas (e.g., citizen science). Users should have the option of checking more than one interest area.

Application Portal / Tool Referral

Referral from an Applications portal or tool is useful in divining preferred spatial resolution (higher is usually better) and data latency (more up to date is better). Furthermore, if the *specific* application can be divined from the referral, the dataset ranking can be even more precise.

Application Keyword Heuristic

Appearance of an keyword related to an application, such as "landslide", may be used to identify the most relevant datasets (e.g., rainfall, soil moisture, ...). Indeed, in many cases, the keyword must be translated to a data content keyword in the search itself to acquire that record in the results in the first place, as "landslide" is unlikely to show up in the description of a rainfall dataset.

Educational Portal / Tool Referral

Referral from an educational portal or tool in general suggests that the Ease of Use heuristic be brought into play. Specific educational portals, e.g., related to global change, may further suggest which datasets are most relevant.

Jargon Keyword Heuristic

Normally an albatross around the necks of EOSDIS search providers, use of jargon in the search terms *suggests* a higher level of experience with EOSDIS data. Examples are platform and instrument acronyms, processing levels, data formats...

Implementation

As should be apparent from the above discussion, there is a rich set of information and heuristics that could be employed to dramatically improve relevancy ranking of data collections. On the one hand, it presents a daunting prospect for actually implementing them given perennial thin funding of data systems. On the other hand, it may be possible to enlist people outside the core EOSDIS development team to do it. This would require constructing the relevancy ranking component of the CMR in such a way that other authorized developers could add what amounts to sorting algorithms. This likely involves a simple API, and ideally a test environment that would not interfere with the core development (such as the Innovation Lab proposed by a working group.) Furthermore, this general area may draw the interest of researchers in data science, who could perhaps bring more sophisticated and powerful methods of relevancy ranking to EOSDIS. (A member of the GES DISC has already expressed interest in something like this for a Master's thesis project; it may also be a good post-graduate internship project.) The CMR relevancy ranking component should be designed to permit (or at least not preclude) this possibility.